I. INTRODUCTION

Brain tumor segmentation from Magnetic Resonance Imaging (MRI) is a critical task in clinical oncology, essential for accurate diagnosis, treatment planning, and monitoring disease progression. Fully Convolutional Neural Networks (FCNNs), particularly the U-Net architecture [2], have become the standard methodology due to their exceptional capability to integrate local and global contextual information. However, optimizing these models for volumetric medical data, such as the BraTS dataset [1], remains challenging. A key hyperparameter influencing both computational efficiency and model generalization is the input patch dimension. This research investigates the non-trivial relationship between input patch size and segmentation performance, aiming to identify the optimal configuration that maximizes the Dice Similarity Coefficient (DSC).

II. OBJECTIVE

The objective of this research is to develop and optimize a robust U-Net segmentation model [2], integrated with advanced image processing for noise mitigation, to enhance segmentation precision. The study focuses on evaluating model performance using the Dice Similarity Coefficient (DSC), analyzing the effect of input patch dimensions on model Generalization capability [2], and proposing implementable improvements for the refined segmentation pipeline.

III. METHODOLOGY

The overall pipeline, from multimodal data acquisition to quantitative evaluation, is structured to isolate and assess the effect of the input patch dimension on the volumetric segmentation task.

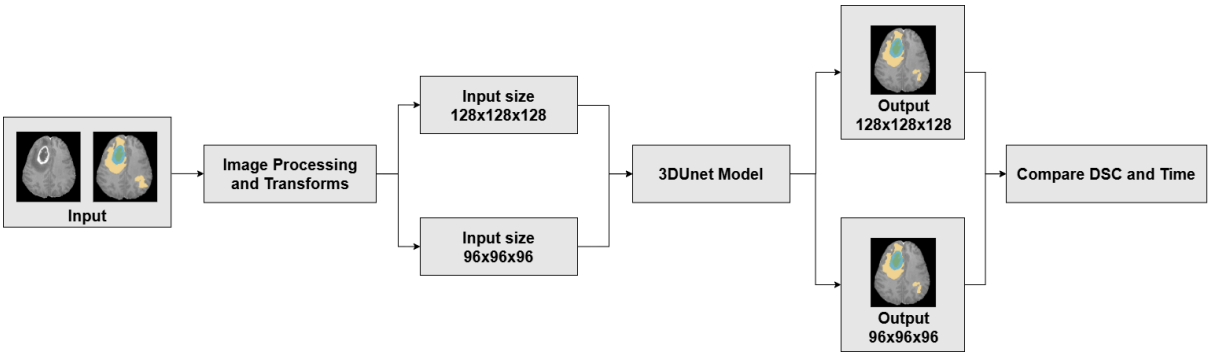


Figure 1: Overall Workflow Block Diagram

The workflow commences with the Input of four multimodal MRI sequences (T1, T1-CE, T2, and FLAIR) along with their corresponding Ground Truth Label Maps. These are passed into the Image Processing and Transforms stage.

The processed data is then utilized to generate two distinct experimental groups based on the input patch dimension: $128 \times 128 \times 128$ voxels and $96 \times 96 \times 96$ voxels. Both inputs feed into the 3D U-Net Model, producing corresponding Output segmentations of $128 \times 128 \times 128$ and $96 \times 96 \times 96$ voxels, respectively. The final step is the Comparative Evaluation of these outputs using the Dice Similarity Coefficient (DSC) and Inference Time metrics.

A. Dataset

The segmentation task utilizes the BraTS 2020 dataset [1], which consists of multimodal volumetric Magnetic Resonance Imaging (MRI) scans of gliomas. This dataset is standardized for evaluating brain tumor segmentation models [1]. Key statistics and technical specifications for the dataset are detailed in Table 2, while the ground truth label mapping is provided in Table 3.

The dataset provides four key sequences (modalities) for each patient, essential for comprehensive tumor characterization: T1-weighted (T1), T1-weighted with contrast enhancement (T1-CE), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR). The training process specifically targets the volumetric segmentation of three distinct tumor sub-regions: the Necrotic Core and Non-Enhancing Tumor (NCR/NET), the Peritumoral Edema (ED), and the Enhancing Tumor (ET), which are the fundamental targets for clinical relevance and ground truth labels [1].

General Dataset Statistics

Characteristic	**Detail**
Total Samples	494 volumes
Training Set Size	369 volumes
Validation Set Size	125 volumes
Input Modalities	4 sequences per patient
Voxel Shape (Input)	$240 \times 240 \times 155$
Voxel Size	$1.0 \times 1.0 \times 1.0^3$
Data Type (Raw Header)	uint8 $96 \times 96 \times 96$
Data Type (Processing)	float64 (typical)

Table 2: Key statistics and technical specifications of the BraTS 2020 dataset.

Ground Truth Label Mapping

The BraTS dataset uses specific labels (0, 1, 2, 4) that are mapped to three binary channels for multi-label segmentation using the ConvertToMultiChannelBasedOnBratsClassesd transform [2].

BraTS Original Label	Region Description	Target Segmentation Channel (Output)
Label 0	Background / Healthy Tissue	Not part of any target channel
Label 1	Necrotic Core (NCR)	Part of Tumor Core (TC) and Whole Tumor (WT)
Label 2	Peritumoral Edema (ED)	Part of Whole Tumor (WT) only
Label 4	Enhancing Tumor (ET)	Part of Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET)
Target TC	Tumor Core (NCR + ET)	Channel 1
Target WT	Whole Tumor (NCR + ED + ET)	Channel 2
Target ET	Enhancing Tumor (ET)	Channel 3

Table 3: Mapping of original BraTS labels to the three target segmentation channels.

B. Image processing and Transform

The data processing pipeline is initiated by loading the raw multimodal MRI volumes and their corresponding ground truth label maps using the LoadImaged command. Subsequently, the ConvertToMultiChannelBasedOnBratsClassesd transform is applied to convert the original BraTS classification (labels 0, 1, 2, 4) into a three-channel tensor, explicitly representing the three key tumor sub-regions: the Tumor Core (TC), the Whole Tumor (WT), and the Enhancing Tumor (ET) [1, 2].

The first essential volumetric preparation step is foreground extraction using CropForegroundd. This transform identifies the bounding box containing the non-zero (foreground) intensity regions within the source image volume. It effectively removes large regions of unnecessary background noise and significantly increases the density of the crucial Region of Interest (ROI) relative to the overall volume, which aids in mitigating class imbalance bias during subsequent training [2].

To ensure consistent input dimensions across all samples for batch processing, the SpatialPadd transform is applied. This operation zero-pads the image and label volumes to a uniform spatial_size defined by config.roi_size using a 'constant' mode. By enforcing a

standardized size, this step guarantees compatibility with the fixed input shape requirements of the 3D U-Net architecture.

Intensity standardization is critical for optimal neural network convergence. The NormalizeIntensityd transform scales the image intensity values by performing channel-wise normalization. Crucially, this normalization process is specifically applied only to non-zero voxels to prevent distortion caused by normalizing against the abundance of zero-valued background voxels, ensuring that feature extraction operates within a stable range [2].

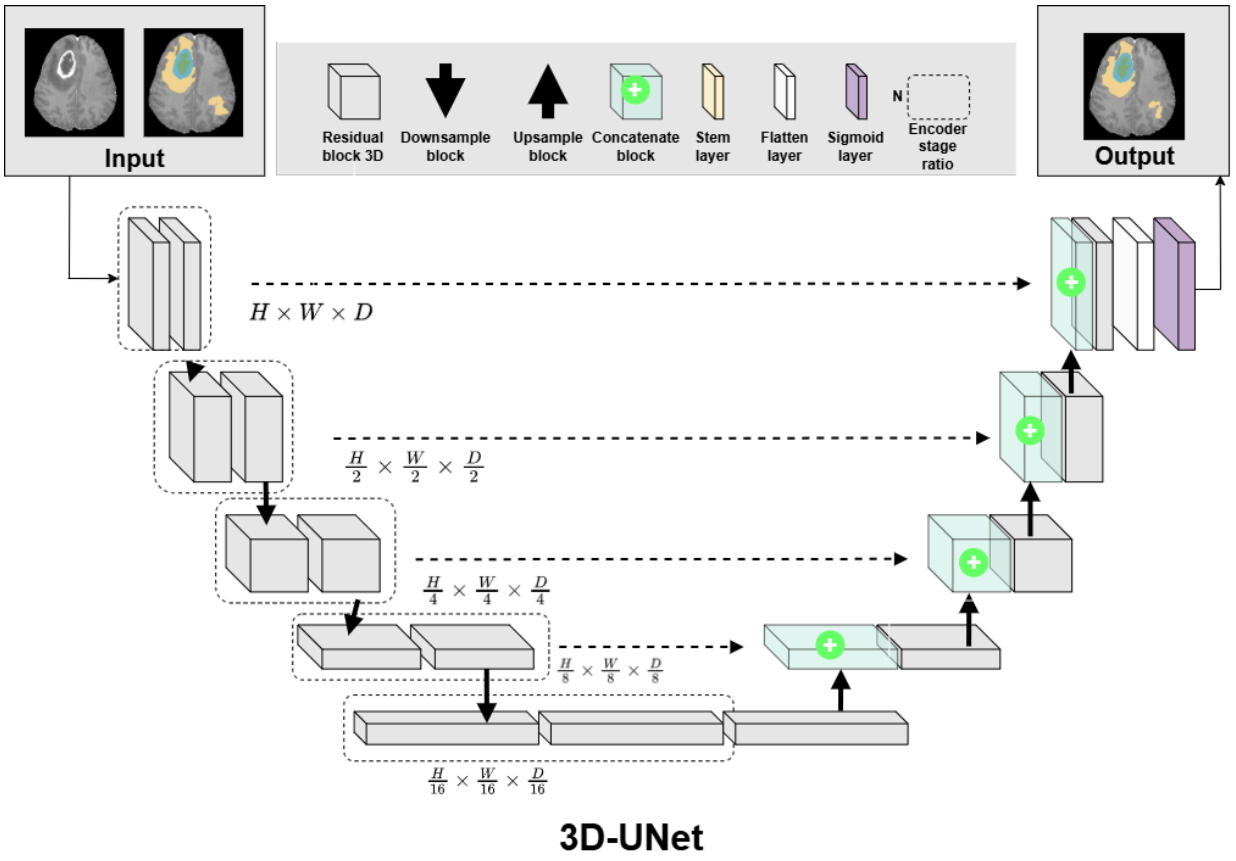
For the training regimen, the RandSpatialCropd technique is employed for patch extraction. This augmentation randomly selects a cubic patch of the predefined size (config.roi_size) from the preprocessed volume. By setting random_size=False, the output patch size remains fixed, which is vital for the controlled experimental comparison of input dimensions.

Geometric data augmentation is introduced to enhance model robustness and generalization against variations in anatomical orientation [4, 5]. The RandFlipd transform applies random flipping along the three spatial axes (0, 1, and 2, corresponding to the depth, height, and width axes, respectively) with a probability ($p = 2$) for each axis. This makes the model invariant to the orientation of the brain scan.

The RandRotate90d transform is utilized to further enrich the geometric variability of the training data. This augmentation applies random 90-degree rotations to the image and label volumes with a probability ($p = 2$), up to a maximum of ($K = 3$) rotations (corresponding to 90, 180, and 270 degrees). This helps the model generalize across various rotational presentations of the tumor within the 3D space [4, 5]. The final step in the pipeline is the conversion of the processed volumes into PyTorch tensors using ToTensord, preparing the data structure for input into the 3D U-Net model.

C. 3D U-Net Architecture and Implementation

The research employed the widely recognized 3D U-Net architecture [2] to perform the volumetric segmentation. The U-Net is selected for its robust capability to capture both fine-grained local details via skip connections and broad global context via the contracting path [2]. This "U-shaped" design forms the core of the model.



3D-UNet

To isolate the impact of the input dimension, all other critical hyperparameters were kept constant across all experiments. The methodology centered on training two distinct groups based on their input Region of Interest (ROI) patch size: the $96 \times 96 \times 96$ Voxel group and the $128 \times 128 \times 128$ Voxel group.

IV. EVALUATIONS

A. Dice Similarity Coefficient (DSC)

DSC is the main quantitative metric used to measure the spatial overlap between the model's prediction (P) and the Ground Truth segmentation (G). It ranges from 0 (no overlap) to 1 (perfect overlap). DSC is widely adopted in medical segmentation due to its robustness against severe Class Imbalance issues [1, 2, 6].

$$DSC = \frac{2\times|P\cap G|}{|P|+|G|}$$

Where $|P \cap G|$ is the volume of overlapping predicted and actual segments (True Positives).

B. Generalization Gap Analysis

Generalization ability was assessed by monitoring the difference between Training Loss and Validation Loss. If the Validation Loss significantly exceeds the Training Loss, it is interpreted as evidence of severe Model Overfitting [2].

VI. RESULTS

The experimental comparison between the two input patch dimensions is quantitatively summarized in Table 1. The results reveal a clear and consistent superiority in segmentation performance and inference efficiency for the smaller input dimension [2].

Metric	**Input 128×128×128 Voxel**	**Input 96×96×96 Voxel**
Val Mean Dice	0.7926	**0.8171**
Val Loss	0.2225	**0.1986**
Train Loss	**0.1595**	0.2078
Dice TC	0.7610	**0.8069**
Dice WT	0.8654	**0.8759**
Dice ET	0.7515	**0.7687**
Time (sec) per picture	1.43	**1.35**

Table 1: Comparison of U-Net performance on the BraTS 2020 dataset across two input patch dimensions [2].

This data confirms that the $96 \times 96 \times 96$ Voxel input achieved a superior Mean Dice Score and demonstrated better generalization (lower Validation Loss), while also providing a reduction in inference time [2]. The significantly lower Training Loss for the 128^3 The model, coupled with its higher Validation Loss, strongly supports the hypothesis of severe overfitting.

VII. DISCUSSION

The experimental results demonstrating the superior performance of the $96 \times 96 \times 96$ Voxel patch size over the $128 \times 128 \times 128$ Voxel patch size, Mean Dice Score: 0.8171, which surpassed the 0.7926 achieved by the larger patch. For the BraTS 2020 segmentation task [1] necessitates a critical re-evaluation of input dimension selection in 3D U-Net architectures [2].

The lower Dice Score observed with the larger 128^3 patch size is primarily attributed to Severe Model Overfitting and a detrimental increase in Class Imbalance [2]. The proportional increase in background voxels within the larger patch significantly dilutes the concentration of the crucial tumor voxels (Region of Interest, ROI). This dilution forces the model to emphasize learning the abundant background class, ultimately impairing its ability to accurately delineate the boundaries of the minority tumor classes (Necrotic, Non-Enhancing, and Enhancing regions).

Conversely, the $96 \times 96 \times 96$ patch size achieved an Optimal Trade-off between two critical segmentation requirements: Context Coverage, which maintained sufficient global and local context necessary for the U-Net architecture to function effectively via its skip connections [2], and ROI Density, which maximized the density of the tumor features relative to the patch size. This finding underscores that for tasks involving small, sparse, and irregular ROIs, the input size must be meticulously tuned as a hyperparameter to ensure optimal Feature Generalization [2].

VIII. CONCLUSION

The selection of the input patch size is a non-trivial factor that directly governs the generalization capability and predictive accuracy of the U-Net model in 3D medical image segmentation [2]. This study decisively concludes that for the BraTS 2020 dataset [1], the $96 \times 96 \times 96$ Voxel dimension represents the Optimal Trade-off size. This size successfully balanced the preservation of tumor feature density and necessary contextual information, resulting in demonstrably superior Model Generalization and the highest recorded Dice Score performance [2], surpassing the performance of the larger 128^3 input.

REFERENCES

1. U. Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," *arXiv preprint arXiv:2107.02314*, 2021.
2. F. Isensee et al., "nnU-Net for Brain Tumor Segmentation," *arXiv preprint arXiv:2011.00848*, 2020.
3. A. Hatamizadeh et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," *arXiv preprint arXiv:2201.01266*, 2022.
4. L. N. Smith, "A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 Learning Rate, Batch Size, Momentum, and Weight Decay," *arXiv preprint arXiv:1803.09820*, 2018.
5. J. J. Luke, R. Joseph, and M. Balaji, "Impact of Image Size on Accuracy and Generalization of Convolutional Neural Networks," *International Journal of Research and Analytical Reviews (IJRAR)*, vol. 6, no. 1, Feb. 2019. (E-ISSN 2348-1269, P-ISSN 2349-5138)
6. K. H. Zou et al., "Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index," *Acad Radiol.*, vol. 11, no. 2, pp. 178–189, 2004.